Data Stream Analisys and Data Mining With Storm

Gregor Majcen, Miha Zidar

***Abstract*—Twitter Storm is a powerfull distributed real time data processing solution, with a wide range of usage. In this paper, we are going to take a look at how we can utilize Twitter Storms power for data mining on streams of data. The main focus of this paper is real time data processing and online data mining.**

***Index Terms*—online learning, continuous data, data mining, distributed systems, horizontal scaling, batch processing**

I. Introduction

TODAY we're generating more information per second than ever before, and the amount of data produced is only increasing over time. Data on its own is not that useful for us, unless we can extract information from it and the speed of gathering that information is becomming more and more valuable. This is where the real time data processing comes in. Big companies like Twitter, Groupon, spider.io and others, are using Twitter Storm to provide a better user experience.

In the last few years data processing has come a long way with services like MapReduce, Amazon EMR, Hadoop, and related techologies. All of these were made to handle massive amounts of data, and they do that very affectively. But lately their weakness is showing in their lack of real time processing. Now speed is starting to be ever more important, to get ahead of competition. This is where Storm comes in. Unlike other solutions that work in "batch" mode, Storm is made so it can handle data streams. With that it's possible to get calculations done on tupples of data and have the results back in matter of seconds, insted of hours or even days. One valuable ability to have, with ever increasing data streams, is scalability. To be more precise, horizontal scalability, where if the data stream grows, we can simply add more nodes withought having to increase the speed of each individual node, and Twitter Storm prodvides that by running on top of Apache Zookeeper cluster.

We will begin by explaining what Storm is, how it works, and what it is used for. Then we shell slowly narow down all the Storm uses to the ones that become useful in datamining. We will show how a simple storm demo project, and discuss how these concepts can be applied to machine learning. At the end we will look at a practical machine learning algorithm running on Storm.

January 3, 2013

II. Twitter Storm

Storm is a distributed and fault-tolerant realtime computation system. It provides a high abstraction layer with witch we can run complex computations on a cluster of computer. Because it runs on top of Zookeeper and has a good messageing system, it provides a good alternative to managing your own cluster with queues and workes.

It can be used for stream processing, processing messages, updating databases, updating online machine learning models in real time. Other uses also include continuous computation, doing a continuous query on data streams and streaming out the results to users as they are computed, and for distributed RPC.

A. Storm structure

Befor we can start anything, we need to learn Storms terminology.

- spout - this is basically the imput for the entire system, it is the point (or multiple points) where the data streams are connected, so that it generate outputs that bolts can read.
- bolt - a simple operating unit, that recieves and processes data from sprouts or bolts and possibly generates an output stream for other bolts
- topology - a complete Storm structior simmilar to a never ending process. It is composed of spouts and bolts and data streams between them. **??**

B. Usage

- Stream processing -
- Distributed RPC
- Continous computation

C. Simple Example

```
builder.setSpout("RedditPostsReader", new RawRed

builder.setBolt("FilterPostString", new filterBo
builder.setBolt("", new ()).shuffleGrouping("");
```

Explanation:

III. Online machine learning

A. Introduction

IV. Machine learning with Storm

A. preprocessing

B. classification

C. learning

V. Conclusion

Acknowledgement

The authors would like to thank Matjaž Kukar, PhD Assistant Professor.

REFERENCES

[1] J. S. Vitter, *Random Sampling with a Reservoir*. Brown University, 1985.

[2] N. Littlestone, *Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm*. University of California, 1988

[3] P. Zhao, R. Jing, *Online AUC Maximization* School of Computer Engineering, Nanyang Technological Unoversoty & Deparment of Computer Science and Engineering, Michigan State University

[4] R. Kessl, *Parallel algorithms for mining of frequent itemsets* The Faculty of Electrical Engineering, Czech Technical University in Prague

[5] Hanley, James A. and McNeil, Barbara J. *The meaning and use of the area under of receiver operating characteristic (roc) curve.* 1982.

Gregor Majcen 63070199

Miha Zidar 63060317